Coupling Vocabularies and Data Structures: Lessons from LOINC™

Roberto A. Rocha, M.D., Ph.D. and Stanley M. Huff, M.D.
Department of Medical Informatics,
University of Utah School of Medicine, Salt Lake City, Utah and
3M Health Information Systems, Murray, Utah

Using LOINC's data model for laboratory test result names as a starting point, an extended model is presented, coupled to a more complete vocabulary model. Justifications for this approach are obtained from a matching experiment that attempts to identify SNOMED terms that correspond to the various components of the LOINC names. Some limitations of LOINC's current vocabulary model, exposed during the matching process, are discussed.

INTRODUCTION

The widespread use of clinical laboratory information systems (CLIS) has motivated the development of message interchange standards, such as Health Level 7 (HL7)[1] and the American Society for Testing and Materials (ASTM) E1238-94.[2] These standards facilitate the electronic transmission of laboratory results between a CLIS and its clients, through a detailed definition of the structure of the exchanged messages.[3] The utilization of these standards has become a requirement for commercial CLIS.

Despite the usefulness of standards describing the structure of the exchanged messages, the content of these messages remained arbitrary.[4] Recognizing the importance of this problem and the existence of partial solutions at best,[5] a group of researchers has created a set of universal names and numeric identifiers for the vast majority of laboratory test results.[4] These test result names are part of the "Logical Observation Identifier Names and Codes" (LOINC) database. The LOINC database is a publicly available coding system created to fully identify concepts that are exchanged between systems using HL7 or ASTM messages.[4]

In the domain of clinical laboratory, the scope of LOINC is limited to test result names, i.e., it does not include names for test order names.[4] The test result names are structured as six complementary segments, each segment describing one or more characteristics of the fully specified name (Figure 1). Some of the more complex segments have sub-components that are individualized using pre-defined delimiting characters (Figure 1). The contents of these various segments are created following naming conventions that specify allowable abbreviations, word order, preferred synonymous forms, punctuation, and case, among others (Figure 1).

METHODS

Acknowledging the usefulness of the model proposed by LOINC for identifying laboratory test result names, 3M Health Information Systems decided to adopt this model for its clinical vocabulary server, called "3M Data Dictionary™" (DD). The hierarchical structure of the LOINC segments was converted into a template structure expressed in Abstract Syntax Notation One (ASN.1).[6] The contents of the various segments and subsegments of the test result names were also converted into subordinate templates, until discrete domains were identified. During this conversion, definitions of the contents of these domains were also created.

The objective of this conversion was to extend the model proposed by LOINC and obtain a model not only capable of identifying all distinct test result names, but also capable of identifying the discrete concepts that, when combined, correspond to the complete meaning of these names. This approach ensured that the internal model adopted by the DD was fully compatible but not limited to the LOINC model, enabling the decomposition of the test result names using unambiguous representations.[7]

The categorization of the discrete domains and their constituent concepts provided the opportunity to express these concepts using terms (surface forms) other than those chosen by LOINC. This functionality is one of the design premises of the DD, enabling translations of the LOINC names to other languages and also to different coding schemes. These translations are crucial to dynamically interface with systems that do not adopt LOINC names and identifiers, to customize the display of test results, and to support queries against the data using alternate surface forms.[7]

In the interest of determining if the model derived from LOINC would improve common interface installation processes such as vocabulary mapping, the original LOINC test result names (version 1.0e)[8] were translated into SNOMED International (SNOMED)[9] codes. Initially, each LOINC name was decomposed into its segments and subsegments, generating a series of files that roughly corresponded to the discrete domains of the new model. All the abbreviated forms found in these files were expanded using the information obtained from the LOINC Users' Guide.[8] Next, the content of these files was

0195—4210/96/$5.00

Test result name structure:

[[Name] . [Subname] . [Sub-subname] ^ [Challenge] ^ [Adjustments] ^ [Person]] tab [Property] tab [Time Aspect] tab [Sample Type] tab [Scale] tab [Method]

Examples:

1. [[AZTREONAM]] tab [SUSC] tab [PT] tab [ISLT+SER] tab [SQ] tab [SBT]
2. [[ALPHA AMYLASE] . [PANCREATIC]] tab [CCNC] tab [PT] tab [SER/PLAS] tab [QN]
3. [[CORTISOL] ^ [1.5H POST DOSE U/KG INSULIN IV]] tab [MCNC] tab [PT] tab [PLAS] tab [QN]
4. [[HLA-DQ LITTLE W4] ^ [ ] ^ [ ] ^ [DONOR]] tab [ACNC] tab [PT] tab [BLD] tab [SQ]

Figure 1 - LOINC test result name structure with four examples.

submitted to a lexical matching process; the details of the matching process are described elsewhere.[10]

The output of the matching process was manually reviewed by both authors, and each match was classified as "exact," "partial" (narrower or broader), or "no-match." Taking into account the presence of discrete ("atomic") concepts in SNOMED, including many generic modifiers, the reviewers were allowed to select more than one SNOMED term to try to represent a single LOINC segment (or subsegment) term.

After the review, an automated process was used to reassemble the LOINC test result names using the SNOMED translations. This same process computed a "match index" for each translated name, based on weights assigned to each segment and subsegment of the LOINC structure (Table 1). Exact matches added the weight of its corresponding segment or subsegment to an intermediary coefficient, while partial matches added only half of the weight to it. For example, if a translated name had only a partial match for the analyte name and an exact match for the sample type, its coefficient would be equal to 4 ((4+2)+2); if the original LOINC name had all six segments defined, its coefficient would be equal to 13 (4+2+2+2+2+1). In this example, the match index would be 0.31 (4÷13).

Table 1 - Weights assigned to each segment and subsegment of the LOINC test result names.

Segment or subsegment	*Weight*
1. Analyte/Component	9
1.1 Name and modifier	6
1.1.1 Name	4
1.1.2 Subname	1
1.1.3 Sub-subname	1
1.2 Challenge information	1
1.3 Adjustments/corrections	1
1.4 Person	1
2. Kind of Property	2
3. Time Aspect	2
4. System/Sample Type	2
5. Type of Scale	2
6. Type of Method	1

RESULTS

Figure 2 (next page) contains the DD model for test result names derived from the LOINC model. Whenever appropriate, the labels used to identify the various templates and attributes of the DD model were adapted from the names used to describe the LOINC segments and subsegments.

Figure 3 shows the frequency distribution of the matching indexes calculated for the LOINC test result names expressed using SNOMED concepts. Figure 4 presents the average matching index for each class of LOINC test result names, as defined in the LOINC Users' Guide. Figure 5 presents the percentage of each match category for each LOINC segment and subsegment defined in Table 1. Table 2 contains some examples of the LOINC-SNOMED matches.

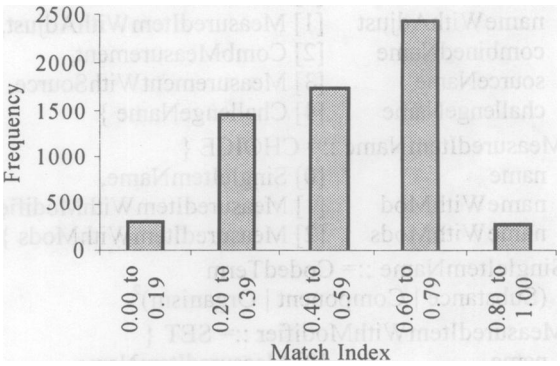

Figure 3 - Frequency distribution of the Match Index.

DISCUSSION

The conversion of the LOINC structure for laboratory test result names was a straightforward process. The resulting data model, despite its apparent complexity, is fully compatible with the original structure proposed by LOINC, while providing the opportunity to make use of a more robust vocabulary model. In other words, instead of being limited to the surface forms adopted by LOINC, the DD model can presumably make use of any synonymous form in any language or coding

```
LOINC DEFINITIONS IMPLICIT TAGS ::=
BEGIN
LoincName ::= SET {
  measuredItem        [0] MeasuredItem,
  propertyOfMeasure   [1] PropertyObserved,
  timing              [2] MeasurementTiming,
  specimenSources     [3] SpecimenSources,
  precision           [4] MeasurementPrecision,
  method              [5] MeasurementMethod OPT }
PropertyObserved ::= CodedTerm
  (ACNC | AREA | CCNC | COLOR)[1]
MeasurementTiming ::= CodedTerm
  (T12H | T24H | T2H | PT)[1]
MeasurementPrecision ::= CodedTerm
  (QL | QN | SQ)[1]
MeasurementMethod ::= CodedTerm
  (CF | CIE | DNA-PROBE | MANUAL-COUNT)[1]
SpecimenSources ::= CHOICE {
  specimen            [0] Specimen,
  multipleSource      [1] MultipleSource,
  eitherSource        [2] EitherSource }
Specimen ::= CodedTerm
  (AMN | CSF | SER | STL | UR)[1]
MultipleSource ::= CodedTerm
  (UR-AND-SER | CSF-AND-SER)[1]
EitherSource ::= CodedTerm
  (SER-OR-PLAS | SER-OR-PLAS-OR-BLD)[1]
MeasuredItem ::= CHOICE {
  name                [0] MeasuredItemName,
  nameWithAdjust      [1] MeasuredItemWithAdjust,
  combinedName        [2] CombMeasurement,
  sourceName          [3] MeasurementWithSource,
  challengeName       [4] ChallengeName }
MeasuredItemName ::= CHOICE {
  name                [0] SingleItemName,
  nameWithMod         [1] MeasuredItemWithModifier,
  nameWithMods        [2] MeasuredItemWithMods }
SingleItemName ::= CodedTerm
  (Substance | Component | Organism)[2]
MeasuredItemWithModifier ::= SET {
  name                [0] MeasuredItemName,
  subName             [1] AnalyteSubName }
MeasuredItemWithMods ::= SET {
  name                [0] MeasuredItemName,
  subName             [1] AnalyteSubName,
  subSubName          [2] AnalyteSubSubName }
MeasuredItemWithAdjust ::= SET {
  measuredItem        [0] MeasuredItemName,
  adjustment          [1] Adjustment OPTIONAL }
Adjustment ::= CodedTerm (PH74)[1]
CombMeasurement ::= SET OF MeasuredItemName
MeasurementWithSource ::= SET {
  name                [0] MeasuredItemName,
  source              [1] SourceName }
SourceName ::= SET {
  nameWithMods        [0] MeasuredItemName,
  analyteSource       [1] AnalyteSource }
ChallengeName ::= SET {
  nameWithMods        [0] MeasuredItemName,
  challengeInfo       [1] ChallengeInfo }
Substance ::= CodedTerm
  (NormalBodySubstance | AbnormBodySubstance)[2]
Component ::= CodedTerm
  (Cell | CellFragments | AggregateOrDeposit)[2]
Organism ::= CodedTerm
  (Bacteria | Ricketsia | Virus | Parasite)[2]
AnalyteSubName ::= CodedTerm
  (Fractionation | Conjugation | AntibodySubtype)[2]
ChallengeInfo ::= SET {
  timeDelay           [0] TimeDelay OPT,
  challengeTime       [1] ChallengeTime,
  substanceAmount     [2] SubstanceAmount OPT,
  substanceOrActivity [3] SubstanceOrActivity,
  substanceRoute      [4] SubstanceRoute OPT }
ChallengeTime ::= CodedTerm (POST | PRE)[1]
TimeDelay ::= CHOICE {
  codedTimeDelay      [0] CodedTimeDelay,
  numericTimeDelay    [1] NumericTimeDelay }
CodedTimeDelay ::= CodedTerm
  (Baseline | Peak | Trough)[1]
NumericTimeDelay ::= SET {
  number              [0] Number,
  timeUnit            [1] TimeUnit }
Number ::= Decimal
TimeUnit ::= CodedTerm
  (Second | Minute | Hour | Day)[1]
AnalyteSubSubName ::= CodedTerm
  (MB | PartialPressure)[1]
SubstanceAmount ::= SET {
  number              [0] Number,
  volumeOrMassUnit    [1] VolumeOrMassUnit }
VolumeOrMassUnit ::= CodedTerm
  (MassUnit | VolumeUnit)[2]
SubstanceOrActivity ::= CodedTerm
  (PhysiologicActivity | Substance | Treatment)[2]
SubstanceRoute ::= CodedTerm
  (PO | SC | IV | IM | IA | ID)[1]
AnalyteSource ::= CodedTerm
  (BPU | CONTROL | DONOR | PATIENT)[1]
END -- Loinc DEFINITIONS
```

Figure 2 - DD model for test result names expressed in ASN.1. Note that some domain definitions have been omitted because of space limitations. [(1)]Sample of a domain defined by LOINC, [(2)]Sample of a domain defined only in this model.

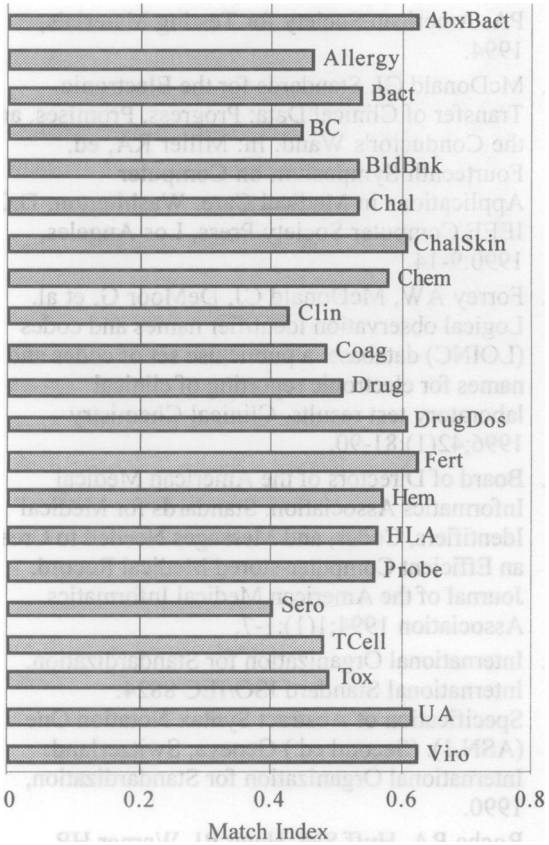

Figure 4 - Average match index per LOINC class.

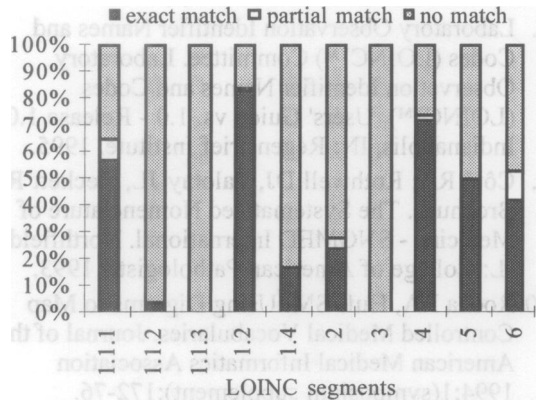

Figure 5 - Percentage of "exact," "partial," and "no matches" per LOINC segment.

scheme. This may not have much influence on the content of the messages exchanged between a CLIS and its clients, since it is likely that codes will be transferred instead of free text. However, this added functionality provides a very useful environment for setting up translation tables between the CLIS master vocabulary and any number of client vocabularies.

In the context of vocabulary mapping, the absence of a mechanism to substitute synonyms and lexical variants is the weakest point of the overall LOINC design. This problem is aggravated by the limitations imposed by the adopted naming conventions. For example, the current LOINC names have no mechanism of representing subscripted and superscripted characters, and the utilization of the keyword "little" to markup segments that should be in lower case has not been correctly implemented. In regard to these two issues, the notation adopted by SNOMED is clearly better. The widespread use of arbitrary abbreviations represents another important limitation of LOINC's vocabulary model, since these abbreviations need to be expanded before lexical matching processes can be successfully used.

Table 2 - Sample of LOINC-SNOMED matches ("M" Match, "e" exact, "p" partial, "n" no match)

LOINC segment	*LOINC surface form*	*M*	*SNOMED surface form*
1.1.1	AZTREONAM	e	[C-53720] Aztreonam
2	SUSC	e	[F-00490] Susceptibility
3	PT	n	
4	ISLT+SER	p	[T-C2500] Serum
5	SQ	n	
6	SBT	e	[P3-55624] Serum inhibitory titer test
1.1.1	CORTISOL	e	[F-B2860] Cortisol
1.2	1.5H POST DOSE U/KG INSULIN IV	p	[G-4004] After + [C-A2200] Insulin preparation, NOS + [G-D101] Intravenous route
2	MCNC	n	
3	PT	n	
4	PLAS	e	[T-C2400] Plasma
5	QN	e	[G-D310] Quantitative

It is clear from the results obtained that SNOMED is not capable of providing complete translations for the fully specified test result names used by LOINC. The plausible explanation for this fact has to do with the scope of these vocabularies. LOINC test result names represent a very detailed and domain-specific vocabulary. SNOMED, on the other hand, represents a much broader vocabulary that spans numerous domains besides clinical pathology. However, the axial structure of SNOMED, in addition to its rich collection of discrete concepts and modifiers, made it the best candidate vocabulary for this experiment. The results demonstrate that segments of the LOINC name that make use of "ordinary" concepts,

particularly the analyte name and the specimen type, are indeed well represented in SNOMED.

Perhaps the most important premise that justifies the utilization of a detailed data model to support vocabulary mapping is the ability to perform aggregations and decompositions of concepts, i.e., the ability to handle "one-to-many," "many-to-one," and "many-to-many" mappings.[7] All three cases were present in this experiment, even after decomposing the LOINC names into segments and subsegments. The least expected cases were aggregate laboratory test names found in the procedures axis of SNOMED, such as "Calcium excretion, 2-hour collection, fasting, urine." In these cases, the SNOMED terms were decomposed before being mapped to the different segments and subsegments of the LOINC names.

CONCLUSION

Vocabulary mapping is certainly one of the most time-consuming steps when interfacing systems. The initiative of the LOINC committee to create a publicly available database with fully specified clinical observation names is a very important step that helps to decrease the complexity of this task. The LOINC database is more than a simple collection of terms, since it is built using a well-defined structure, where names result from the aggregation of concepts from several discrete domains. The central goal of this study was to once more demonstrate the advantages of coupling a vocabulary with a formal underlying data structure.

References

1. Health Level Seven. Health Level Seven Standard Version 2.2: An Application Protocol for Electronic Data Exchange in Healthcare Environments. Ann Arbor, MI: Health Level Seven, Inc., 1994.
2. ASTM E1238-94. Standard Specification for Transferring Clinical Observation Between Independent Computer Systems. Philadelphia, PA: American Society for Testing Materials, 1994.
3. McDonald CJ. Standards for the Electronic Transfer of Clinical Data: Progress, Promises, and the Conductor's Wand. In: Miller RA, ed. Fourteenth Symposium on Computer Applications in Medical Care. Washington, D.C.: IEEE Computer Society Press, Los Angeles, 1990:9-14.
4. Forrey AW, McDonald CJ, DeMoor G, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clinical Chemistry 1996;42(1):81-90.
5. Board of Directors of the American Medical Informatics Association. Standards for Medical Identifiers, Codes, and Messages Needed to Create an Efficient Computer-stored Medical Record. Journal of the American Medical Informatics Association 1994;1(1):1-7.
6. International Organization for Standardization. International Standard ISO/IEC 8824: Specification of Abstract Syntax Notation One (ASN.1). (Second ed.) Geneva, Switzerland: International Organization for Standardization, 1990.
7. Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a Controlled Medical Vocabulary Server: The VOSER Project. Computers and Biomedical Research 1994;27(6):472-507.
8. Laboratory Observation Identifier Names and Codes (LOINC™) Committee. Laboratory Observation Identifier Names and Codes (LOINC™) Users' Guide vs. 1.0 - Release 1.0e. Indianapolis, IN: Regenstrief Institute, 1995.
9. Côté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L. The Systematized Nomenclature of Medicine - SNOMED International. Northfield, IL: College of American Pathologists, 1993.
10. Rocha RA, Huff SM. Using Digrams to Map Controlled Medical Vocabularies. Journal of the American Medical Informatics Association 1994;1(symposium supplement):172-76.